



NASA and The Semantic Web

Naveen Ashish
NASA Ames Research Center

NASA



- Missions
 - Exploration
 - Science
 - Aeronautics
- IT Research & Development at NASA
 - Focus on supercomputing, networking and intelligent systems
- Enabling IT technologies for NASA missions
- NASA FAA research in Air Traffic Management



Semantic Web at NASA

- NASA does not do 'fundamental' semantic web research
 - Development of ontology languages, semantic web tools etc.
- Applications of SW technology to NASA mission needs
- Scattered across various NASA centers such as Ames, JPL, JSC etc.



Various Projects and Efforts

- Collaborative Systems
 - Science, Accident Investigation
- Managing and Accessing Scientific Information and Knowledge
- Enterprise Knowledge Management
- Information and Knowledge Dissemination
 - Weather data etc.
- Decision Support and Situational Awareness Systems
 - System Wide Information Management for Aerospace
- Scientific Discovery
 - Earth, Environmental Science etc.

Still have only taken "baby steps" in the direction of the Semantic Web

SemanticOrganizer



Collaborative Systems

- The Semantic Organizer
- Collaborative Knowledge Management System
 - Supports distributed NASA teams
 - Teams of scientists, engineers, accident investigators ...
 - Customizable, semantically structured information repository
- A large Semantic Web application at NASA
 - 500 users
 - Over 45,000 information nodes
 - Connected by over 150,000 links
- Based on shared ontologies



Repository

- Semantically structured information repository
- Common access point for all work products
- Upload variety of information
 - Documents, data images, video, audio, spreadsheet
-
- Software and systems can access information via XML API

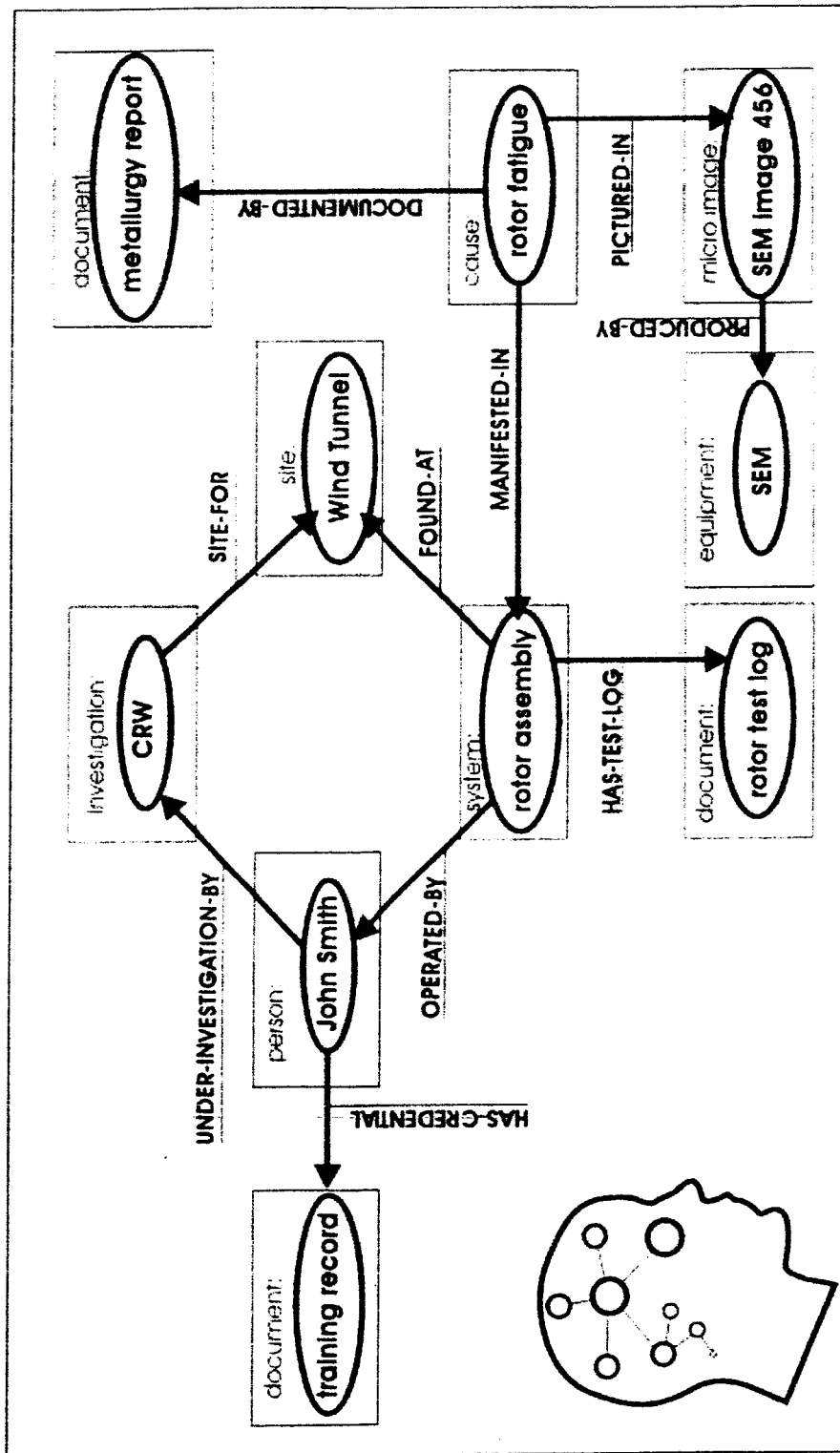


Unique NASA Requirements

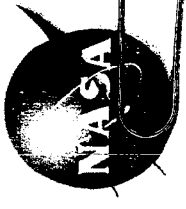
- Several document and collaborative tools in market
- NASA distinctive requirements
 - Sharing of heterogeneous technical data
 - Detailed descriptive metadata
 - Multi-dim correlation, dependency tracking
 - Evidential reasoning
 - Experimentation
 - Instrument-based data production
 - Security and access control
 - Historical record maintenance



SemanticOrganizer

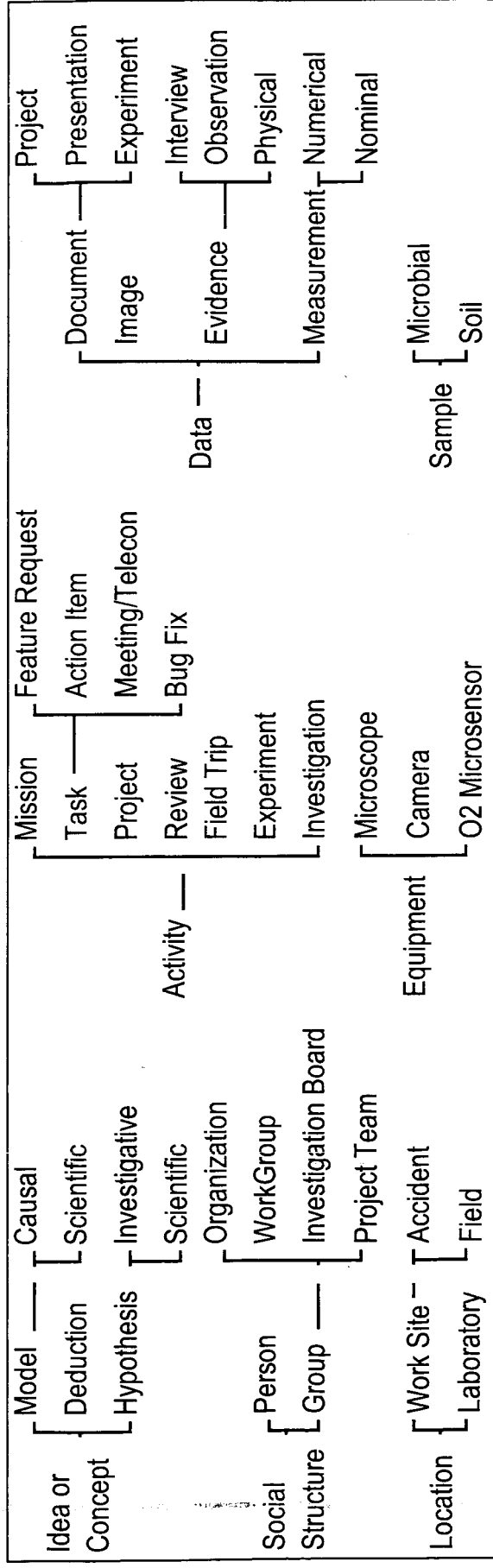


■ <data items and links figure>



Master Ontology

- Master ontology
- Custom developed representation language
 - Equivalent expressive power to RDFS



Implementation

- System built in Java
- Ontology and classes stored in MySQL
- Inference component built on top of Jess





create new item instance

search for items

icon identifies item type

modify item

Current Item

ScienceOrganizer: Baja May June 2001 - Microsoft Internet Explorer

File Edit View Favorites Tools Help

ScienceOrganizer: An Information Sharing Tool For Scientific Project Teams

New Item View Links Edit Links Search Home

May-June 2001 (open all | close all)

Associated Document (3 Trip Documents)

- Baja Report 7/2001
- Brad's Objectives for May-June Trip
- Strawman for Focus Group Samples

Destination (1 Study Areas)

- Baja

Has Experiment (1 Experiments)

- In Situ Diel 6-4-01

Has Logistics Described In (1 Trip Documents)

- Strawman for Focus Group Samples

Has Objectives (1 Trip Documents)

- Brad's Objectives for May-June Trip

Has Participant (15 Persons)

- Has Photodocumentation (99 Images)
- Baja June 2001 Photos
- Samples Collected (19 Samples)
- Measurement Taken (7 Measurements)
- Trip For (1 Projects)
- EMERG

Field Trip: May-June 2001

Item ID# 17540 updated 2001/10/18 01:08PM PDT by Bebout, Brad

Send this Item's web address via Email

Modify Permissions Delete Duplicate Put in a Folder Logout Help

Study Area Baja

End Date 06/12/2001

Start Date 05/30/2001

Trip Objectives First Focus Group Trip

Notes Conduct joint diel experiments to establish baseline

Write Permission Bebout, Brad EMERG

Read Permission Bebout, Brad EMERG

Participants

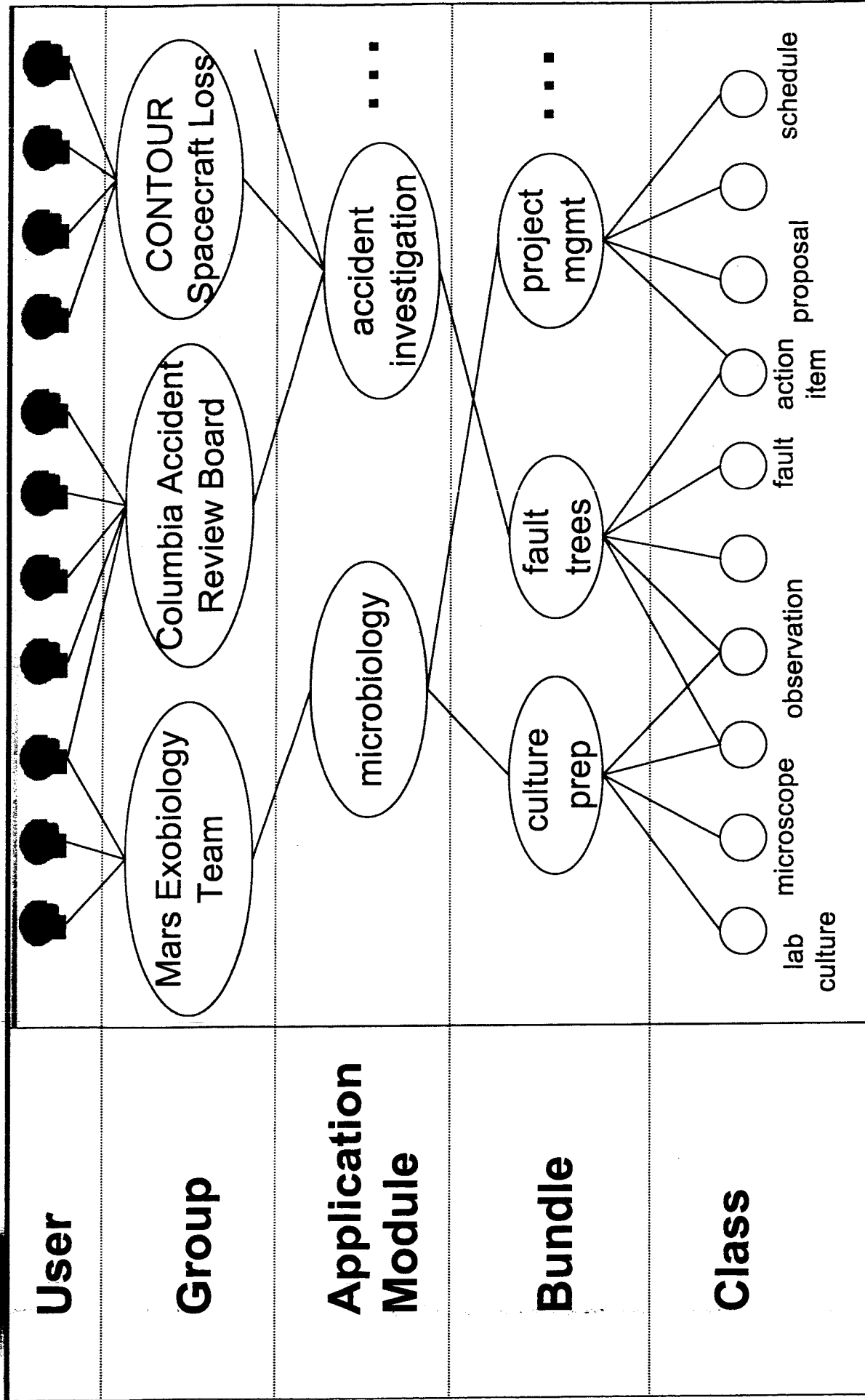
- Bebout, Brad
- Carpenter, Steven
- Des Marais, David
- Dillon, Jesse
- Farmer, Jack

Right side displays metadata for the current repository item being inspected

Left side uses semantic links to display all information related to the repository item shown on the right



Application Customization Mechanism





Applications

- One of the largest NASA Semantic Web applications
 - 500 users, a half-million RDF style triples
 - Over 25 groups (size 2 to 100 people)
 - Ontology has over 350 classes and 1000 relationships
- Scientific applications
 - Distributed science teams
- Field samples
 - *collected-at: , analyzed-by: , imaged-under:*
- Early Microbial Ecosystems Research Group (EMERG)
 - 35 biologists, chemists and geologists
 - 8 institutions



InvestigationOrganizer

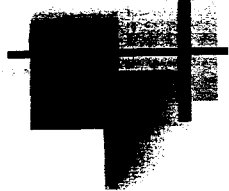
- NASA accidents
 - Determine cause
 - Formulate prevention recommendations
- Information tasks
 - Collect and manage evidence
 - Perform analysis
 - Connect evidence
 - Conduct failure analyses
 - Resolution on accident causal factors
- Distributed NASA teams
 - Scientists, Engineers, Safety personnel
- Various investigations
 - Space Shuttle Columbia, CONTOUR

Lessons Learned



- Network structured storage models present challenges to users
- Need for both 'tight' and 'loose' semantics
- Principled ontology evolution is difficult to sustain
- Navigating a large semantic network is problematic
 - 5000 nodes, 30,000-50,000 semantic connections
- Automated knowledge acquisition is critical

The NASA Taxonomy





NASA Taxonomy

- Enterprise information retrieval
 - With a standard taxonomy in place
- Development
 - Done by Taxonomy Strategies
 - Funded by NASA CIO Office
- Design approach and methodology
 - With the help of subject matter experts
 - Top down
 - Ultimately to help (NASA) scientists and engineers find information

Best Practices



- Industry best practices
 - Hierarchical granularity
 - Polyhierarchy
 - Mapping aliases
 - Existing standards
 - Modularity
- Interviews
 - Over 3 month period
 - 71 interviews over 5 NASA centers
 - Included subject matter experts in unmanned space mission development, mission technology development, engineering configuration management and product data management systems. Also covered managers of IT systems and project content for manned missions



Taxonomy

- <http://nasataxonomy.jpl.nasa.gov>



NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

+ Visit NASA.gov
+ Contact the NASA Curator

- TAXONOMY TOP LEVEL FACETS	+ FAQs	+ NASA METADATA	+ NASA TAXONOMY XML	+ NASA XML PROJECT
-----------------------------	--------	-----------------	---------------------	--------------------

NASA Taxonomy - Top Level Facets

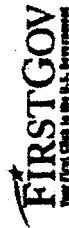
- Access Security Requirements
- Audiences
- Business Purpose
- Competencies
- Content Types
- Industries
- Instruments
- Locations
- Missions and Projects
- Organizations
- Subject Categories

Tips on using the NASA Taxonomy

What is the NASA taxonomy?

The NASA taxonomy is a controlled vocabulary that is designed to populate the NASA metadata core specification.

It is also a means of tagging NASA content so that it can be used and reused in many different contexts.



- + Freedom of Information Act
- + The President's Management Agenda
- + FY2002 Agency Performance Report



Contact the NASA Curator
NASA Official: Jayne Dulra
Last Updated: May 25, 2004



Metadata

- Purpose
 - identify and distinguish resources
 - provide access to resources through search and browsing
 - facilitate access to and use of resources
 - facilitate management of dynamic resources
 - manage the content throughout its lifecycle including archival
- Uses Dublin Core schema as base layer
- NASA specific fields
 - Missions and Projects
 - Industries
 - Competencies
 - Business Purpose
 - Key Words
- <http://nasataxonomy.jpl.nasa.gov/metadata.htm>



In Action: Search and Navigation

- Browse and search
- Seamark from Siderean



NATIONAL AERONAUTICS
AND SPACE ADMINISTRATION

Search Go

219,958 items

by Organization

[NASA Affiliated Institutions](#) 1378
[NASA Centers](#) 76545
[NASA Contractors](#) 10108
[NASA Enterprises](#) 815
[NASA Headquarters](#) 4042
[Other NASA Partners](#) 999

by Subject

[Aeronautics](#) 26532
[Astronautics](#) 31759
[Chemistry and Materials](#) 17086
[Engineering](#) 39631
[Geosciences](#) 30770
[Mathematical and Computer Sciences](#) 13286
[Space Sciences](#) 22685
[4 more](#)

by Missions and Projects

[Aerospace Technology](#) 60
[Biological and Physical Research](#) 68
[Data](#) 140
[Earth Sciences](#) 1497
[Human Exploration and Development](#) 10680
[Planetary Missions](#) 4819
[Space Sciences](#) 9467

by Date

[1940s](#) 111
[1950s](#) 1113
[1960s](#) 10523
[1970s](#) 64244
[1980s](#) 60031
[1990s](#) 67388
[2000s](#) 16452
[4 more](#)

by Competencies

[Business](#) 386
[Engineering](#) 303
[Mission](#) 555
[Science](#) 410
[Technical](#) 218

by Information Type

[Catalogs and Databases](#) 32
[Designs and Specifications](#) 62
[Plans and Agendas](#) 158
[Results and Analyses](#) 260
[Reviews and Lessons Learned](#) 1819
[Status Reports](#) 119
[Technical Reports](#) 229
[6 more](#)

by Collection

[Lessons Learned](#) 1370
[NTRS](#) 213900
[SIRTE](#) 4054
[Webb](#) 634

Search and Navigation



NATIONAL AERONAUTICS
AND SPACE ADMINISTRATION

173 items matching

Text contains manned lunar mission (X)

Refine

by Text search

Go

by Organization
NASA Centers 62
NASA Contractors 14
NASA Headquarters 2

by Subject
Aeronautics 7
Astronautics 75
Chemistry and Materials 4
Engineering 18
Geosciences 3
Life Sciences 12
Space Sciences 28
4 more

by Missions and Projects
Human Exploration and Development... 42
Planetary Missions 36
Space Sciences 4

by Date
1960s 14

Search

Go

Sort by Date Title 1 to 10 of 173 items | next

Mission analysis for the third manned lunar mission

Creator: Yang, T. L. (Bellcomm, Inc., Washington, DC, United States)
Subject: Astronautics
Date: 1969-01 (January)
Collection: NTRS

Manned lunar program options - mission equipment

Creator: Bendersky, C. (Bellcomm, Inc., Washington, DC, United States), Valley, D. R. (Bellcomm, Inc., Washington, DC, United States)
Subject: Ground Support Systems and Facilities (Space)
Date: 1967-01 (January)
Collection: NTRS

Manned lunar program options - mission equipment

Creator: Bendersky, C. (Bellcomm, Inc., Washington, DC, United States), Valley, D. R. (Bellcomm, Inc., Washington, DC, United States)
Subject: Astronautics
Date: 1967-01 (January)
Collection: NTRS

Assessment of manned lunar mission technological support

Creator: Herd, J. H. (Operations Research, Inc., Silver Spring, MD, United States), Olmer, F. J. (Operations Research, Inc., Silver Spring, MD, United States), Parent, S. R. (Operations Research, Inc., Silver Spring, MD, United States)
Subject: Astronautics
Date: 1962-01 (January)
Collection: NTRS

Manned lunar program options mission modes



Near Term Implementations

- The NASA Lessons Learned Knowledge Network
- NASA Engineering Expertise Directories (NEEDs)
- The NASA Enterprise Architecture Group
- NASA Search

SWEET: The Semantic Web of Earth and Environmental Terminology



SWEET



- SWEET is the largest ontology of Earth science concepts
- Special emphasis on improving search for NASA Earth science data resources
 - Atmospheric science, oceanography, geology, etc.
 - Earth Observation System (EOS) produces several Tb/day of data
- Provide a common semantic framework for describing Earth science information and knowledge

Prototype funded by the NASA Earth Science Technology Office



Ontology Design Criteria

-  **Machine readable:** Software must be able to parse readily
-  **Scalable:** Design must be capable of handling very large vocabularies
-  **Orthogonal:** Compound concepts should be decomposed into their component parts, to make it easy to recombine concepts in new ways.
-  **Extendable:** Easily extendable to enable specialized domains to build upon more general ontologies already generated.
-  **Application-independence:** Structure and contents should be based upon the inherent knowledge of the discipline, rather than on how the domain knowledge is used.
-  **Natural language-independence:** Structure should provide a representation of *concepts*, rather than of terms. Synonymous terms (e.g., marine, ocean, sea, oceanography, ocean science) can be indicated as such.
-  **Community involvement:** Community input should guide the development of any ontology.



Global Change Master Directory (GCMD) Keywords as an Ontology?

- Earth science keywords (~1000) represented as a taxonomy. Example:
EarthScience>Oceanography>SeaSurface>SeaSurfaceTemperature
- Dataset-oriented keywords
 - Service, instrument, mission, DataCenter, etc.
- GCMD data providers submitted an additional ~20,000 keywords
 - Many are abstract (climatology, surface, El Nino, EOSDIS)



SWEET Science Ontologies

- Earth Realms
 - Atmosphere, SolidEarth, Ocean, LandSurface, ...
- Physical Properties
 - temperature, composition, area, albedo, ...
- Substances
 - CO₂, water, lava, salt, hydrogen, pollutants, ...
- Living Substances
 - Humans, fish, ...



SWEET Conceptual Ontologies



- Phenomena
 - ElNino, Volcano, Thunderstorm, Deforestation, Terrorism, physical processes (e.g., convection)
 - Each has associated EarthRealms, PhysicalProperties, spatial/temporal extent, etc.
 - Specific instances included
 - e.g., 1997-98 ElNino
- Human Activities
 - Fisheries, IndustrialProcessing, Economics, Public Good



SWEET Numerical Ontologies

- SpatialEntities
 - Extents: country, Antarctica, equator, inlet, ...
 - Relations: above, northOf, ...
- TemporalEntities
 - Extents: duration, century, season, ...
 - Relations: after, before, ...
- Numerics
 - Extents: interval, point, 0, positiveIntegers, ...
 - Relations: lessThan, greaterThan, ...
- Units
 - Extracted from Unidata's UDUnits
 - Added SI prefixes
 - Multiplication of two quantities carries units



Numerical Ontologies (cont.)

- Numeric concepts defined in OWL only through standard XML XSD spec
 - Intervals defined as restrictions on real line
- Numerical relations defined in SWEET
 - lessThan, max, ...
- Cartesian product (multidimensional spaces) added in SWEET
- Numeric ontologies used to define spatial and temporal concepts

Data Ontology



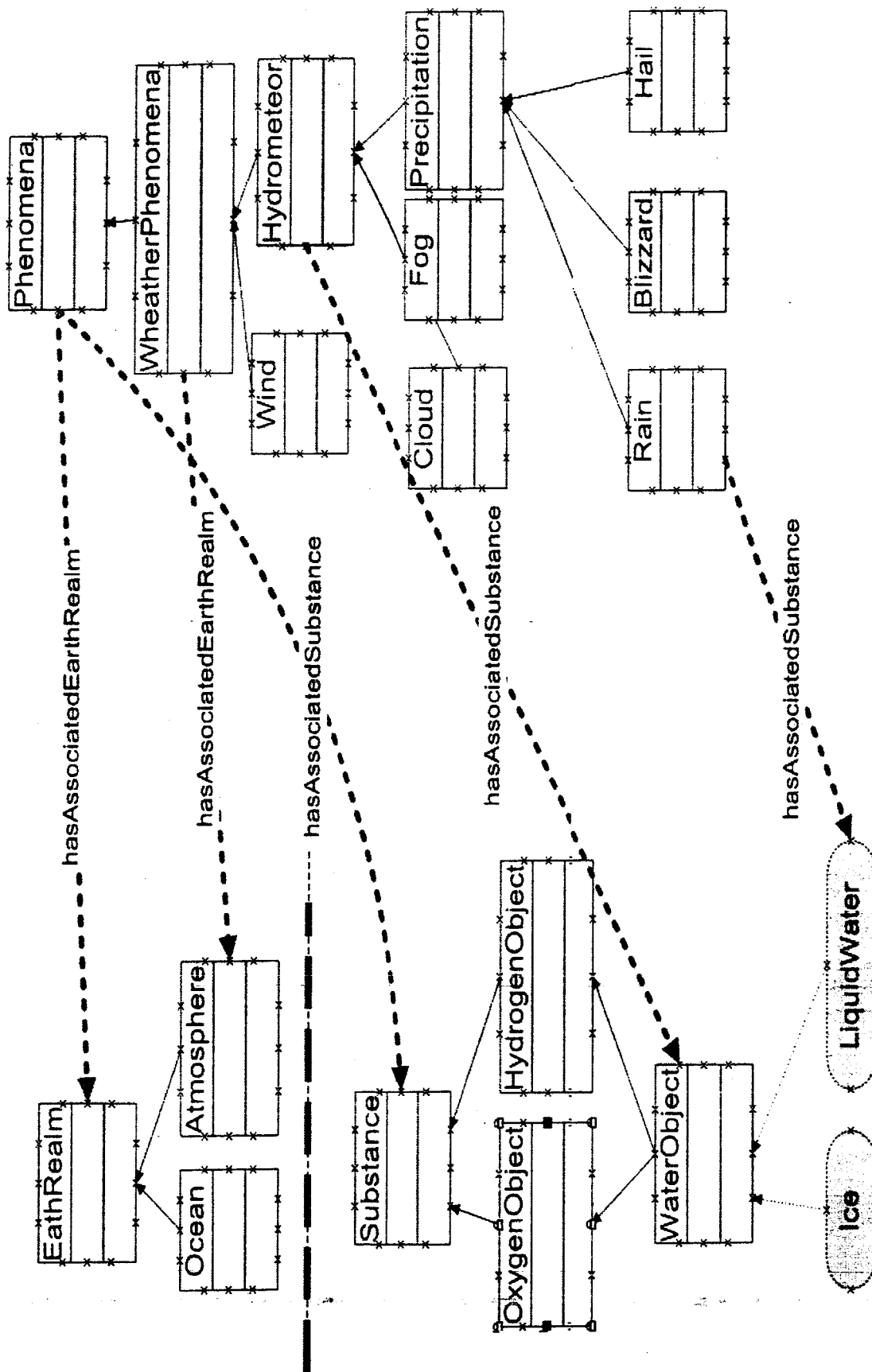
- Data semantics
- Special values
 - Missing, land, sea, ice, etc.
- Parameters
 - Scale factors, offsets, algorithms

Spatial Ontology



- Polygons used to store spatial extents
 - Most gazetteers store only bounding boxes
- Polygons represented natively in Postgres DBMS
- Includes contents of large gazetteers
- Stores spatial attributes (location, population, area, etc.)

Ontology Schematic





Example: Spectral Band

```
<owl:class rdf:ID="VisibleLight">
  <rdfs:subclassOf>
    ElectromagneticRadiation
  </rdfs:subclassOf>
  <rdfs:subclassOf>
    <owl:restriction>
      <owl:onProperty rdf:resource="#Wavelength" />
      <owl:toClass owl:class="Interval400to800" />
    </rdfs:subclassOf>
  </owl:class>
```

- Class "Interval400to800" separately defined on PhysicalQuantity
- Property "lessThan" separately defined on
- "moreEnergetic" is subclass of "lessThan" on ElectromagneticRadiation

Ontologies as a Unifying Knowledge Framework



- Orthogonal concepts
 - Property, substance, space, time, ...
 - "Science reductionism"
- Unifying concepts
 - Phenomena
 - "Science synthesis"

An OWL Compiler for Mathematical Entities?



- Some numerical constructs in an OWL document could be read by a compiler
 - Operators $\langle \text{add} \rangle$, $\langle \text{divide} \rangle$, $\langle \text{min} \rangle$, etc.
 - Numerical values
- Compiler directives

DBMS Storage



- DBMS storage desirable for large ontologies
- Postgres
- Two-way translator
 - Converts DBMS representation to OWL output on demand
 - Imports external OWL files

How Will OWL Tags Get Onto Web Pages?



- 1. Manual insertion:
 - Users insert OWL tags to each technical term on a Web page
 - Requires users to know of the many ontologies/namespaces available, by name
- 2. Automatic (virtual) insertion:
 - Tags inferred from context while the Web pages are scanned and indexed by a robot
 - Tags reside in indexes, not original documents

Clustering/Indexing Tools



- Latent Semantic Analysis
 - A large term-by-term matrix tallies which ontology terms are associated with other ontology terms
 - Enables clustering of multiple meanings of a term
 - e.g. Java as a country, Java as a drink, Java as a language
- Statistical associations
 - Similar to LSA, but heuristic
- Will be incorporated in ESIP Federation Search Tool



Earth Science Markup Language (ESML)

- ESML is an XML extension for *describing* a dataset and an associated library for *reading* it.
- SWEET provides semantics tags to *interpret* data
 - Earth Science terms
 - Units, scale factors, missing values, etc.

Earth Science Modeling Framework (ESMF)



- ESMF is a common framework for large simulation models of the Earth system
- SWEET supports *model* interoperability
 - Earth Science terms
 - Compatibility of model parameterizations, modules

Earth System Grid (ESG)



- ESG is a computational grid for large simulation models of the Earth system
- SWEET supports *grid* interoperability
 - Earth Science terms
 - Grid concepts



Federation Search Tool

- ESIP Federation search tool
 - SWEET looks up search terms in ontology to find alternate terms
 - Union of these terms submitted to search engine
- Version control & lineage
- Metrics
 - Representing outcomes and impacts

Semantics: Shared Understanding of Concepts



- Provides a namespace for scientific terms....plus
- Provides descriptions of how terms relate to one another
- Example tags in markup language:
 - subclass, subproperty, part of, same as, transitive property, cardinality, etc.
- Enables object in "data space" to be associated formally with object in "science concept space"
- "Shared understanding" enables software tools to find "meaning" in resources

Knowledge Reuse



- SWEET is a concept space
- Enables scalable *classification* of Earth science concepts
 - Search engines use Open directory to classify contents of WWW space
- For educational use, SWEET supports navigable discovery of Earth science concepts, such as for an electronic encyclopedia
- Concept space is translatable into other languages/cultures using "sameAs" notions

Contributions of SWEET

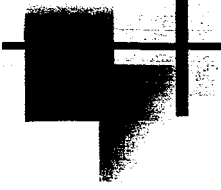


- Improved data discovery without exact keyword matches
- SWEET Earth Science ontologies will be submitted to the OWL libraries

Contacts

- SWEET <http://sweet.jpl.nasa.gov>
- Rob Raskin raskin@jpl.nasa.gov
- Mike Pan mjipan@jpl.nasa.gov



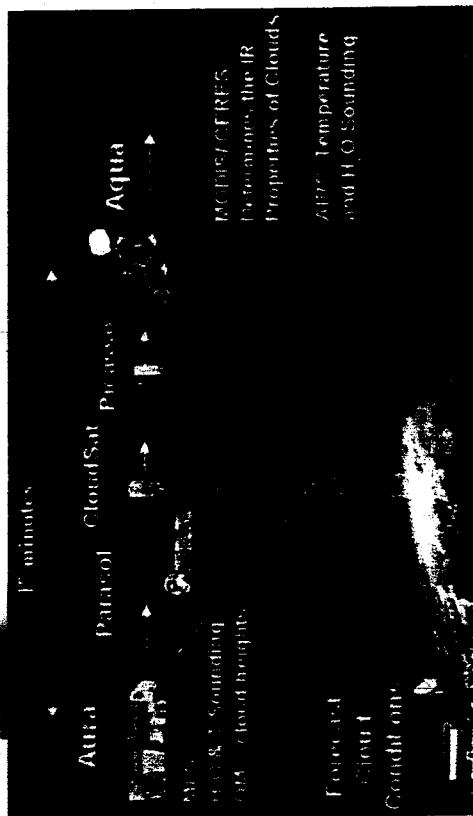


NASA Discovery Systems Project

Project Objective

Create and demonstrate new discovery and analysis technologies, make them easier to use, and extend them to complex problems in massive, distributed, and diverse data enabling scientists and engineers to solve increasingly complex interdisciplinary problems in future data-rich environments.

Discovery Systems Project



Scientists and engineers have a significant need to understand the vast data sources that are being created through various NASA technology and projects. The current process to integrate and analyze data is labor intensive and requires expert knowledge about data formats and archives. Current discovery and analysis tools are fragmented and mainly support a single person working on small, clean data sets in restricted domains. This project will develop and demonstrate technologies to handle the details and provide ubiquitous and seamless access to and integration of increasingly massive and diverse information from distributed sources. We will develop new technology that generates explanatory, exploratory, and predictive models, makes these tools easier to use, and integrate them in interactive, exploratory environments that let scientists and engineers formulate and solve increasingly complex interdisciplinary problems. Broad communities of participants will have easier access to the results of this accelerated discovery process.

Technologies to be included:

- Collaborative exploratory environments and knowledge sharing
- Machine assisted model discovery and refinement
- Machine integration of data based on content
- Distributed data search, access and analysis

Enterprise Requirements



Distributed data search, access and analysis

- Produce customizable data products, data processing and analysis for the wide variety of NASA stakeholders
- Allow seamless multidisciplinary access to and operation on massive, distributed archives of heterogeneous data, models, and data processing algorithms
- Evolvement and archive data systems
- Instruments and platforms need to be integrated with large-scale computing and data systems.
- Data, models, and associated algorithms should be detailed, complete, easily located, catalogued, documented, and organized by content
- Efficiently use communication resources to get maximal value out of data stored in remote and distributed systems

CODES

All

All

All

R, S, U, Y
M, S, Y, U

Machine integration of data based on content

- Data and algorithm interoperability - users do not have to cope with translations and interpolations of the data either across missions or disciplines.
- Data integration - heterogeneous data are automatically registered, reconciled and fused/merged prior to analysis for both real-time and retrospective studies.
- Data validation and annotation - primary and derived data products should include all information necessary for (re)analysis, including explicit representation of source, experimental context, uncertainty, and pedigree

All

All

All

Machine-assisted model discovery and refinement

- Discover and understand complex behavior across vast heterogeneous data sets
- Automated methods to create explanatory, exploratory, and predictive models of complex data
- Automated methods to identify trends and events, track changes, summarize results, and identify information gaps
- Methods to effectively model complexity and covariability of data at different spatial and temporal scales
- Visualize / navigate / explore / mine investigation results faster, on new data types, at higher volumes
- Use system simulations as predictive models in data analysis applications and closed-loop model-prediction-driven targeted data generation or requests
- Methods to improve evaluation and comparison of alternative models

All

All

M, R, S, Y

All

All

Collaborative exploratory environments and knowledge sharing

- Support efficient interdisciplinary research and working groups, e.g., Astrobiology Institute, Instrument science teams, mission investigations, integrated product development teams for aerodynamics
- Increase the value of data and knowledge about the data for multiple users across geographic and time barriers
- Increase diversity of participants across skill levels and reduce organizational barriers to

U, Y, S, R

All

All

All

Enterprise Applications



Distributed data search, access and analysis

- Framework to enable a self-managing network of data repositories, processes, and instruments that can support resource allocation to 100,000 simultaneous end-users and automated algorithms.
- Development and demonstration of virtual data systems that support user-defined computations and intermediate and deferred access, processing, and distribution.
- Methods for usable fine-grain resource access and policy enforcement based on security, trust, and commercial concerns
- Automatic and opportunistic optimization of data analysis methods for resource-effective execution over distributed architectures
- Algorithms and systems for search and analysis across diverse structured and unstructured data including semantic data, models, algorithms, documents, and conversations

Machine integration of data based on content

- Formal process and framework for planning, integrating, and publishing metadata and semantics about a new resource so that new resources can be integrated in 1 week.
- Algorithms and processes to bridge different semantic ontologies and link diverse mission databases in 1 month
- Methods for data validation (quality-control) and annotation that integrate semantic models and external data sources, including explicit representation of source, experimental context, uncertainty and provenance (pedigree)
- Custom data product creation system that automatically integrates and reconciles heterogeneous data and models and republishes a self-annotated product

Machine-assisted model discovery and refinement

- Algorithms to enable requests or collection of new data guided by automated analysis of model predictions and observations
- Methods to reduce massive high-dimensional databases by a factor of 100 while preserving information for task specific investigations.
- Data mining approaches that support discovery and analysis as volume, complexity, and diversity of data/models increases by a factor of 10 over a baseline capability
- Methods to effectively model complexity and covariability of remote sensing and other data sources at different spatial and temporal scales.
- Method to exploit structures uncovered by machine learning algorithms to learn about the problem space in which you are trying to do optimization
- Methods to accurately encode domain expertise to enhance the efficiency, scalability, accuracy, and usability of knowledge discovery algorithms

Collaborative exploratory environments and knowledge sharing

- Task-specific knowledge environments that can be specialized for multi-disciplinary investigations 1 week
- Integrated discovery assistants that assist in problem formulation, solution construction, and task execution, mining and understanding using distributed resources
- Methods to preserve 75% of the utility of PI generated data and knowledge about the data for multiple users across vast geographic distances over a period of 100 years.



Discovery Systems Before/After

Technical Area	Start of Project	After 5 years (In-Guide)
Distributed Data Search Access and Analysis	Answering queries requires specialized knowledge of content, location, and configuration of all relevant data and model resources. Solution construction is manual.	Search queries based on high-level requirements. Solution construction is mostly automated and accessible to users who aren't specialists in all elements.
Machine integration of data / QA	Publish a new resource takes 1-3 years. Assembling a consistent heterogeneous dataset takes 1-3 years. Automated data quality assessment by limits and rules.	Publish a new resource takes 1 week. Assembling a consistent heterogeneous dataset in real-time. Automated data quality assessment by world models and cross-validation.
Machine Assisted Model Discovery and Refinement	Physical models have hidden assumptions and legacy restrictions. Machine learning algorithms are separate from simulations, instrument models, and data manipulation codes.	Prediction and estimation systems integrate models of the data collection instruments, simulation models, observational data formatting and conditioning capabilities. Predictions and estimates with known certainties.
Exploratory environments and collaboration	Co-located interdisciplinary teams jointly visualize multi-dimensional preprocessed data or ensembles of running simulations on wall-sized matrixed displays.	Distributed teams visualize and interact with intelligently combined and presented data from such sources as distributed archives, pipelines, simulations, and instruments in networked environments.

Data Access



Before

- Finding data in distributed databases
 - based on way it was collected
 - (ie a specific instrument at a specific time)

After

- Finding data in distributed databases
 - search by kind of information
 - (ie show me all data on volcanoes in northern hemisphere in the last 30 years)

Data Integration



Before

- Publishing a new resource
 - 1-3 years
- Assembling of consistent heterogeneous datasets
 - 1-3 years

After

- Publishing a new resource
 - 1 week
- Assembling of consistent heterogeneous datasets
 - real-time

Data mining



Before

- El Nino detection and impact on terrestrial systems.
 - Frequency : annual-monthly
 - Resolution : Global 0.5 Deg down to 8km terrestrial and 50km Ocean
 - Relationships: Associative rules between clusters or between raster cells.

After

- El Nino detection and impact on terrestrial systems.
 - Frequency : monthly-daily (or 8-day composite data)
 - Resolution : Global: 0.25km terrestrial and 5km Ocean
 - Relationships: causal relationships between ocean, atmospheric and terrestrial phenomena

Causal relationships



Before

- Causal relationships detectable within a single, clean database
 - NASA funded "data providers" assemble datasets in areas of expertise "WEBSTER" for global terrestrial ecology.
 - Data quality required: heavily massaged and organized
 - Data variety: 1-5 datasets with 1-3 year prepwork

After

- Causal relationships detectable across heterogeneous databases where the relationships are not via the data but via the world
 - Carbon cycle, sun-earth connection, status of complex devices
 - Heterogeneous data, multi-modal samplings, multi spatio-temporal samples.
 - Data quality required: direct feed
 - Data variety: 10 datasets with 1 month prep work

Knowledge Discovery Process



Before

- Iterate data access, mining, and result visualization as separate processes.
- Expertise needed for each step.
- The whole iteration can take months.
 - Some specialty software does discovery at the intersection of COTS with NASA problems: ARC/Info, SAS, Matlab, SPlus and etc.

After

- Integration between access, mining, and visualization
 - explore, mine, and visualize multiple runs in parallel, in real-time
- Real-time data gathering driven by exploration and models
- Reduce expertise barriers needed for each step.
- Extend systems to include use of specialty software where appropriate: ARC/Info, SAS, Matlab plug-ins, SPlus and etc.



Discovery Systems Project

- Draft L1 Milestones -

- **Programmatic**
 - 4Q 2005 Identify challenge problems for the Enterprises, and engage the extramural research community with NASA's high priority technology needs in knowledge access and discovery.
Enterprise workshop, RFI, and NRA reflecting these inputs
- **Distributed data search, access and analysis**
 - 2008 Virtual data product creation system including a combination of existing data, interpolated data, pipeline, running simulation, and running instrument, with various constraints and goals in the query and resource-efficient intelligent execution.
 - 2009 Data product creation system enables novel what-if and predictive question answering by non data-specialists across heterogeneous data and simulations.
- **Semantic modeling and integration**
 - 2009 Machine integration of noisy, sparse, multi-resolution, heterogeneous data for impact analysis.
 - 2009 Automated semantic cross-validation Q&A of data and dealing with the consequences of defects and invalid data. (Finding something in a sensor data stream creates expectations for other, co-varying sensors).
- **Automated knowledge discovery**
 - 2009 Demonstrate a physically-based time-series prediction of dynamical atmospheric and or surface phenomena with hyper-spectral images; identify attractor states; provide associated uncertainties in retrieved measurements and data products.
 - 2009 Demonstrate an architecture for data-efficient simulations that is scalable to enable high-resolution climate predictions without exceeding hardware capability as projected by Moore's law (integrated milestone with ACAT project)
 - 2005 Initial sensor web prototype / systems analysis (integrated milestone with CDS project)
 - 2008 Demonstrate Interactive self-improving simulations driving a simulated sensor web (w/CDS)
- **Exploratory environments and Collaboration**
 - 2008 Demonstrate exploration environments integrating automated discovery technologies and distributed data pipelines with intelligent agents and process capture

Discovery Systems Projects

- Customer/Product Users and Partners -



Enterprise Customers

- Earth Science Enterprise
 - EOSDIS/SEEDS
 - Automated meta-data generation (provenance & content)
 - Algs for distributed, on-line discovery of novel models
 - Science program
 - Tech to extract patterns and models from massive databases of heterogeneous data.
 - Discovery frameworks to discover novel models within complex databases describing complex systems
 - Virtual archives for instruments and platforms
 - Data request generation for sensor-web and prediction systems
 - Discovery and change detection algorithms for onboard science & compression
 - Applications Program (customer)
 - Tech for building reliable decision support systems
 - System for custom data product creation
- Space Science Enterprise
 - Virtual Observatory/observatory missions
 - "Living with a Star"
 - Planetary exploration
 - Astrophysics Data Analysis Program (ADAP)
 - Discovery Data Analysis Program (DDAP)
 - Long-Term Space Astrophysics Program (LTSA)
 - Mars Data Analysis Program (MDA)

Proposed Partners

- Government Agencies
 - NSF (eg. Cyber-Infrastructure Initiative),
 - DARPA (eg. CogSys, EELD, TIA, GENOA), ONR, AFRL
- Academia/Industry
 - Leverage exploratory science environments (eg CACTUS) and Grid initiatives (eg Globus, GriPhyN, EUDatagrid)
 - Major aerospace companies
 - Major universities

- Life Science Enterprise (Biological & Phys Research)
 - Advanced Human Support Technology Program
 - » Technologies for modeling the human body in non-earth environments
 - » Frameworks for sharing and fusing data from studies on multiple physiologic systems in multiple subjects
 - Space Radiation Biology
 - » Technologies to extract patterns from disparate databases on cancer initiation and radiation repair
 - » Discover mechanisms of molecular damage
 - NASA Astrobiology Institute
 - Interdisciplinary exploratory environments
 - Cellular and Macromolecular Biotechnology Program
 - » Technologies to extract patterns from disparate databases on cellular and molecular processes in zero-g
 - Technologies to extract patterns from disparate databases on cellular and molecular processes in zero-g
- Office of Space Flight
 - Spaceport technology
 - » PRACA access and mining
 - Launch and range operations
 - » Integrate data from models and experiments
 - Mission operations
 - » Problem resolution data access
 - Self-sufficiency
 - » Medical and patient information mgmt
- Aerospace
 - SWIM Program, AvSafety and AvSec
 - » Handoff partner for most low-TRL elements
 - ECS as handoff partner for data access, search, discovery, and exploration environments



Discovery Systems Project

- WBS Technology Elements -

- **Distributed data search, access and analysis**
 - Grid based computing and services
 - Information retrieval
 - Databases
 - Planning, execution, agent architecture, multi-agent systems
 - Knowledge representation and ontologies
- **Machine-assisted model discovery and refinement**
 - Information and data fusion
 - Data mining and Machine learning
 - Modeling and simulation languages
- **Exploratory environments and Collaboration**
 - Visualization
 - Human-computer interaction
 - Computer-supported collaborative work
 - Cognitive models of science



Discovery Systems Project

- Major Project Deliverables: DRAFT Milestones -

- **Programmatic**
 - 4Q 2005 Identify challenge problems for the Enterprises, and engage the extramural research community with NASA's high priority technology needs in knowledge access and discovery.
 - Enterprise workshop, RFI, and NRA reflecting these inputs
- **Distributed data search, access and analysis**
 - 2008 Virtual data product creation system including a combination of existing data, interpolated data, pipeline, running simulation, and running instrument, with various constraints and goals in the query and resource-efficient intelligent execution.
 - 2009 Data product creation system enables novel what-if and predictive question answering by non data-specialists across heterogeneous data and simulations.
- **Semantic modeling and integration**
 - 2009 Machine integration of noisy, sparse, multi-resolution, heterogeneous data for impact analysis.
 - 2009 Automated semantic cross-validation Q&A of data and dealing with the consequences of defects and invalid data. (Finding something in a sensor data stream creates expectations for other, co-varying sensors).
- **Machine-assisted model discovery and refinement**
 - 2009 Demonstrate a physically-based time-series prediction of dynamical atmospheric and or surface phenomena with hyper-spectral images; identify attractor states; provide associated uncertainties in retrieved measurements and data products.
 - 2009 Demonstrate an architecture for data-efficient simulations that is scalable to enable high-resolution climate predictions without exceeding hardware capability as projected by Moore's law (integrated milestone with ACAT project)
 - 2005 Initial sensor web prototype / systems analysis (integrated milestone with CDS project)
 - 2008 Demonstrate Interactive self-improving simulations driving a simulated sensor

Ontology Negotiation



Ontology Negotiation

- Allow agents to co-operate
 - Even if based on different ontologies
- Developed protocol
 - Discover ontology conflicts
 - Establish a common basis for communicating
 - Through incremental interpretation, clarification and explanation
- Efforts
 - DARPA Knowledge Sharing Initiative (KSI)
 - Ontolingua
 - KIF

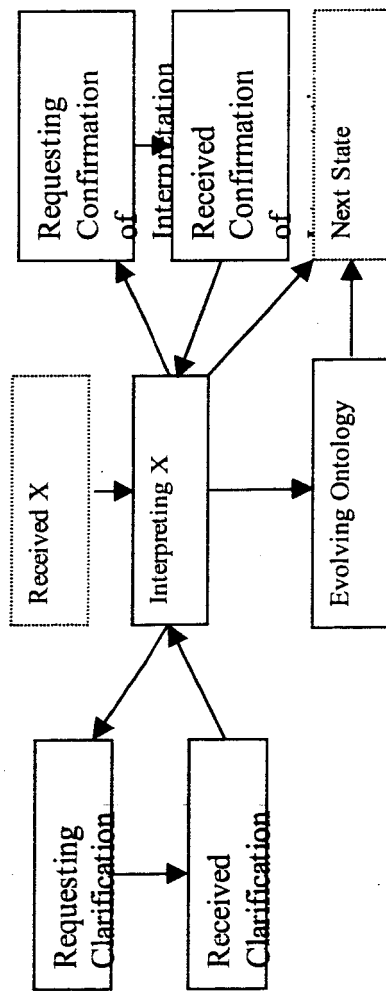
Solutions



- Existing solutions
 - standardization, aggregation, integration, mediation, open ontologies, exchange
- Negotiation



Negotiation Process

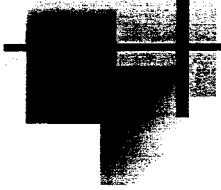


- Interpretation
- Clarification
- Relevance analysis
- Ontology evolution



NASA Scenario

- Mediating between 2 NASA databases
 - NASA GSFC's GCMD
 - NOAA's Wind and Sea archive
- Research on interactions between global warming and industrial demographics
 - Scientists agents
 - Request for clarification



System Wide Information Management

Introduction



- Scenario:
 - Bad weather around airport
 - Landing and take-off suspended for two hours
 - Flights in-flight rerouted and scheduled flights delayed or cancelled
- Passenger inconvenience, financial losses
- Can the situation be handled *efficiently* and *optimally* ?

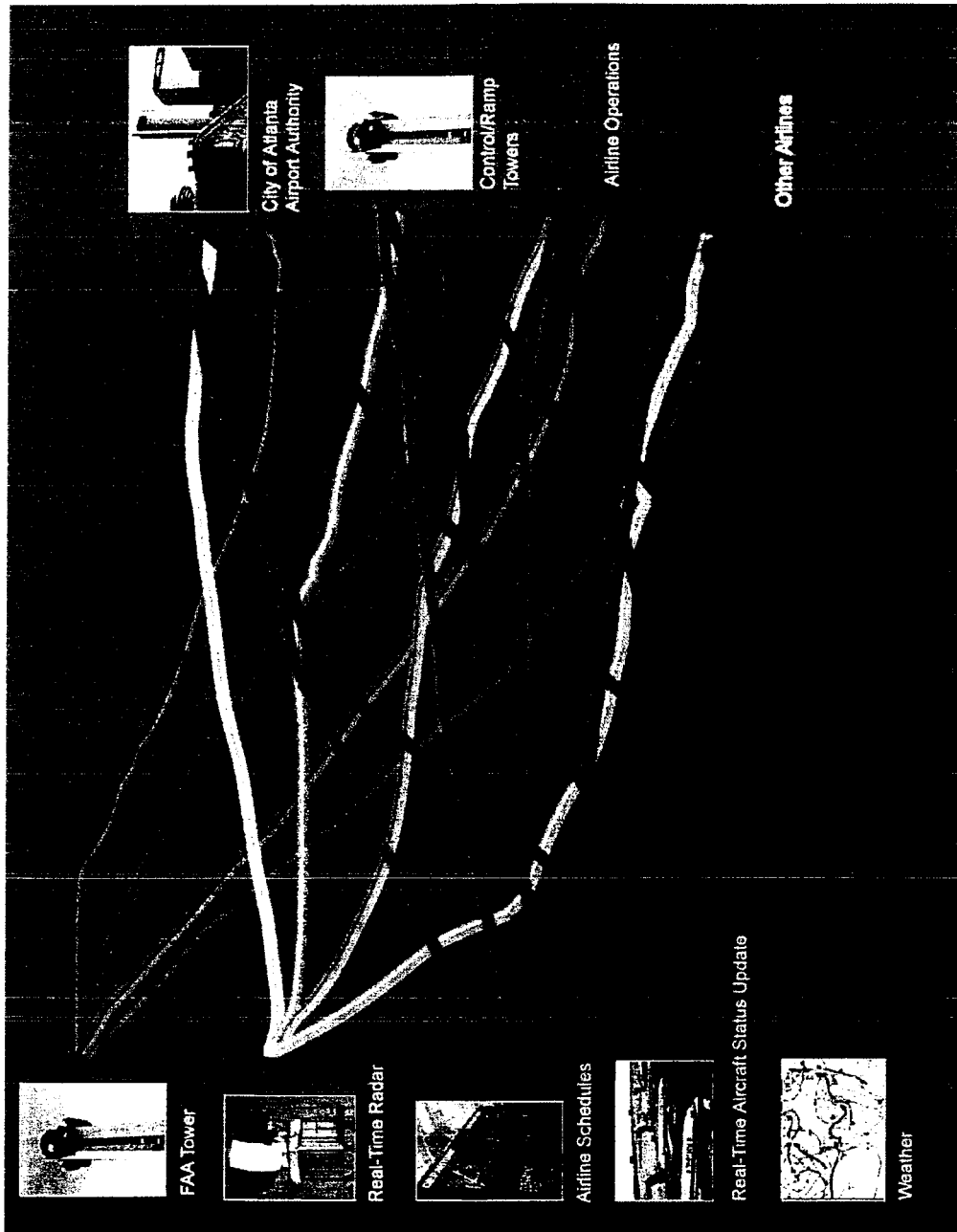
System Wide Information Management



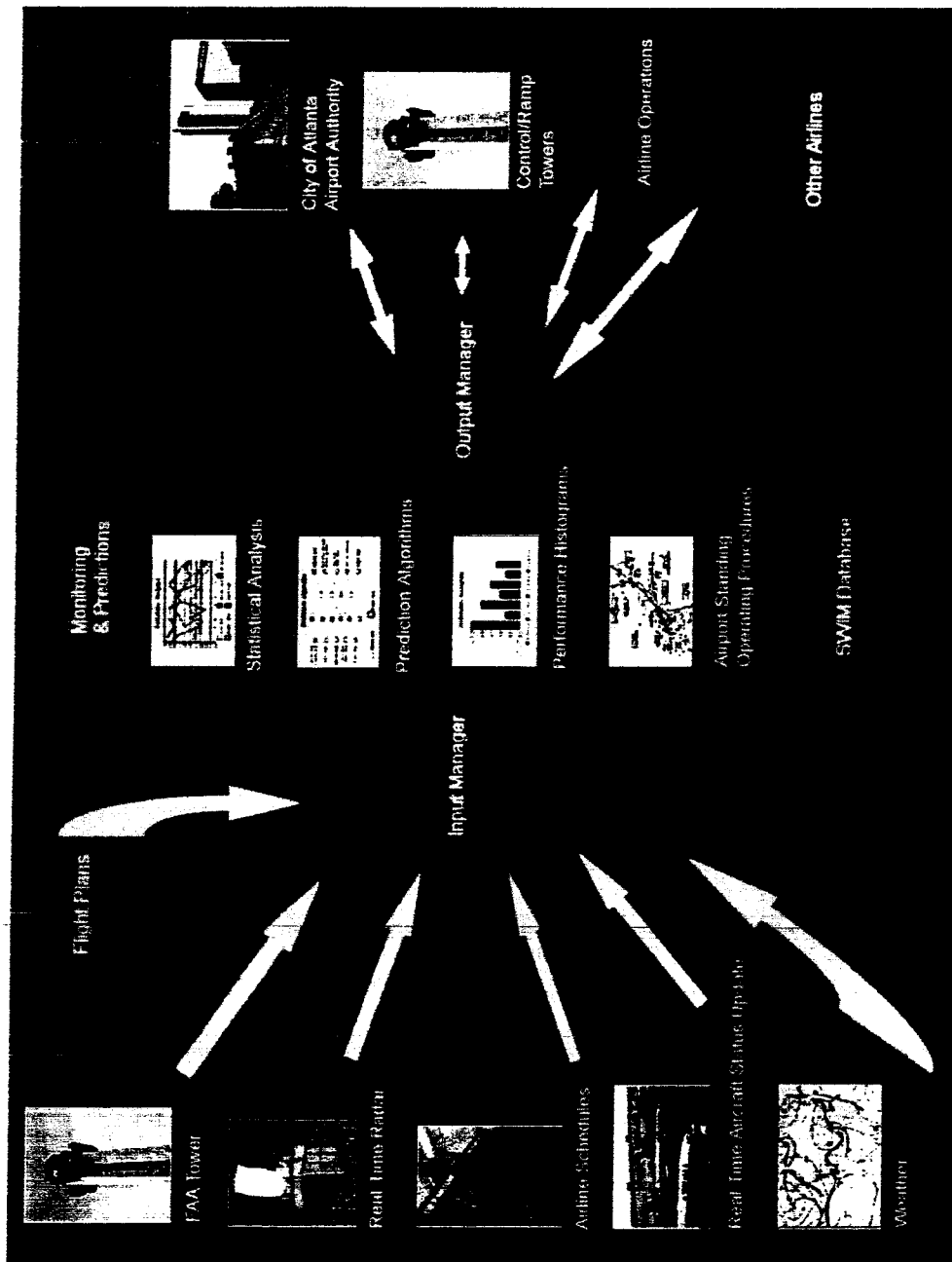
- National Airspace System (NAS)
 - Interconnected network of computer and information sources
- Vision
 - Intelligent agents to aid in decision support
- Decision Support Tools (DSS) use information from multiple heterogeneous sources
- Critical problem is Information Integration !



Present



With SWIM



NAS Information



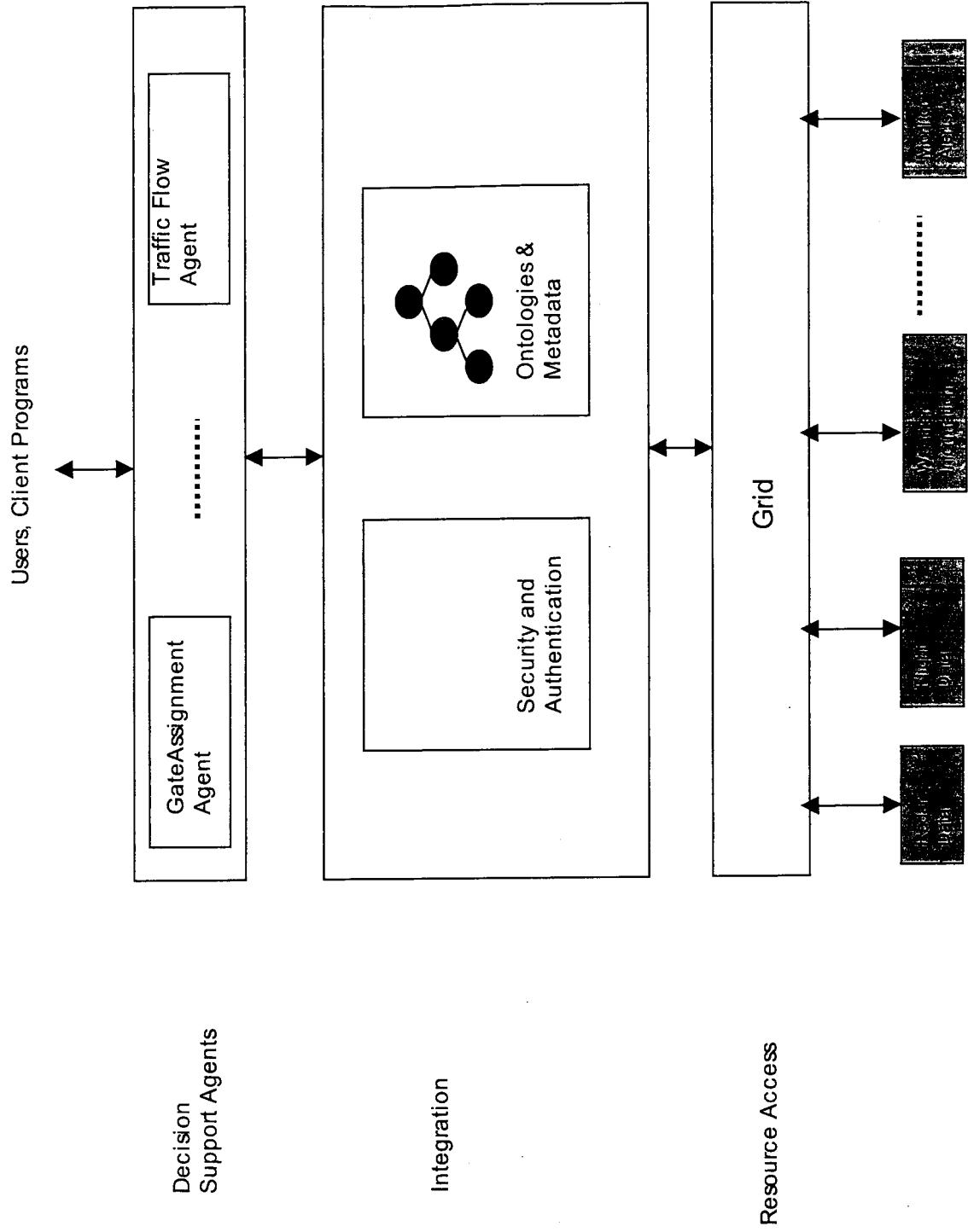
- Information in the NAS comes from a wide *variety* of information sources and is of different kinds
 - Georeferenced information, weather information, hazard information, flight information
- There are different kinds of *systems* providing and accessing information
 - Tower systems, oceanic systems, TFM systems, ...
- Various Categories of DSS Tools
 - Oceanic DSS, Terminal DSS, Enroute DSS,

The Semantic-Web Approach



- Evolved from the information mediation approach
- Key concepts
 - Standard markup languages
 - Standard ontologies
- Can build search and retrieval agents in this environment
- Markup initiatives in the aviation industry
 - AIXM
 - NIXL

Achitecture

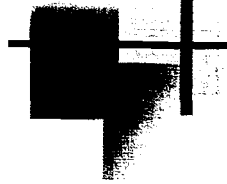




Conclusions

- Introduction to NAS domain
 - DSS tools
 - Information access requirements
 - Nature of information and data sources
- Technology Appropriateness (Initial)
 - Agents
 - Semantic-Web
 - Information retrieval

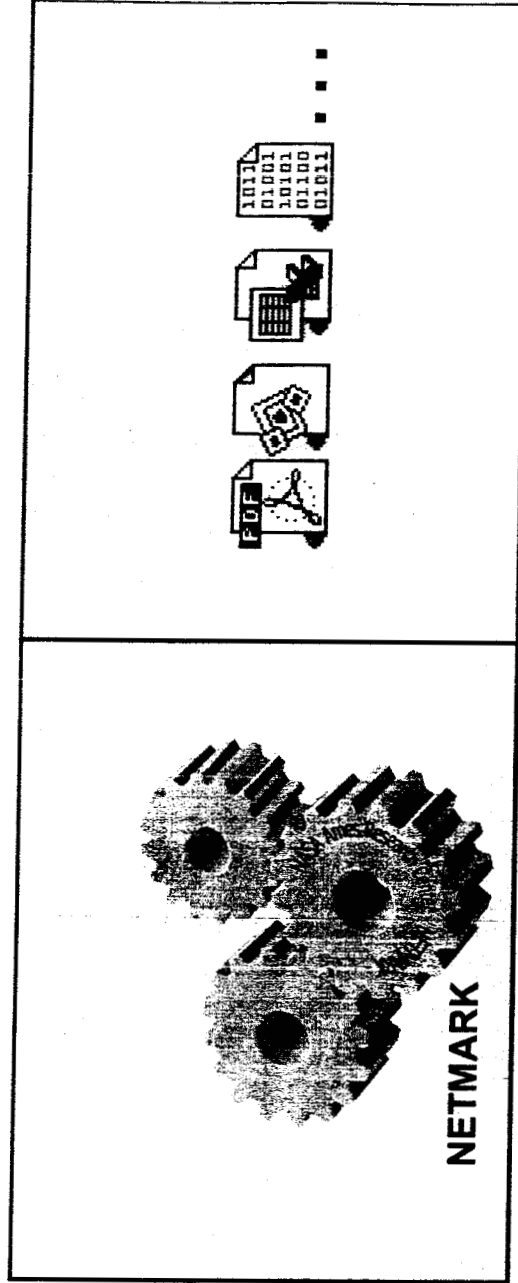
NETMARK: Enterprise Knowledge Management



NETMARK



■ Managing semi-structured data



Load seamlessly into Netmark

Context plus Content search

Regenerate arbitrary documents from arbitrary fragments

to some extent ...garbage in, garbage out.

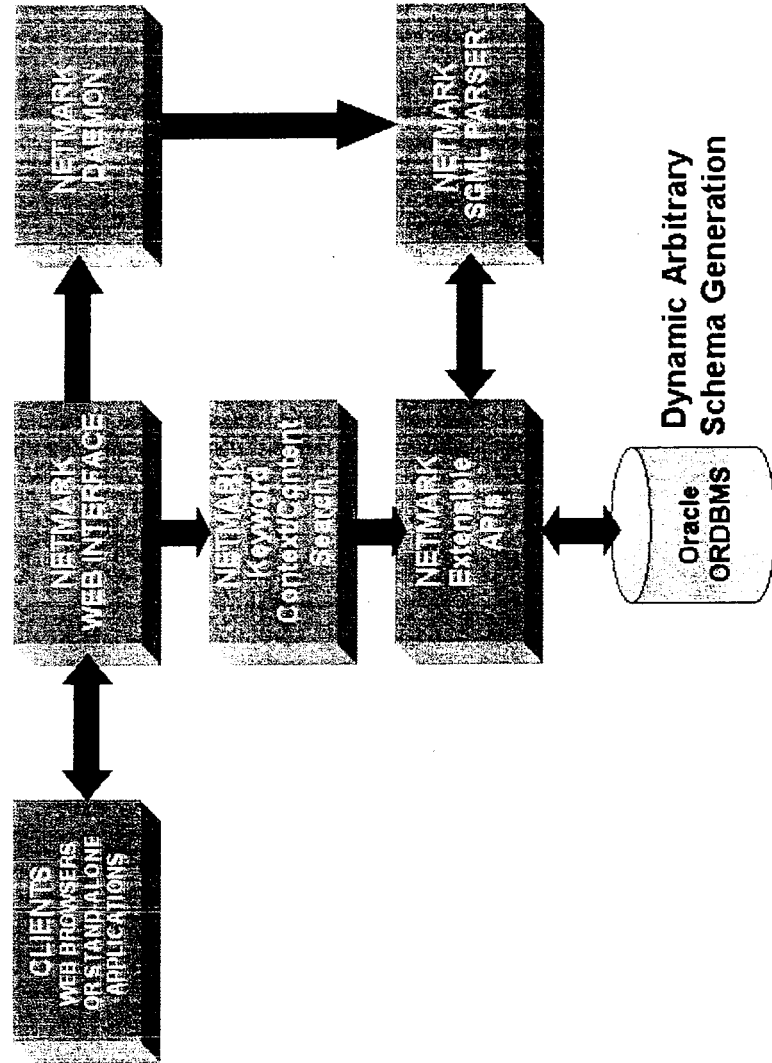


Figure 2: NETMARK Universal Process Flow

Conclusions



- Intelligent information integration and retrieval continue to be key and challenging problems for NASA
 - Science, Aviation, Engineering, Enterprise, ..
- Semantic-web technologies have been/are being successfully applied
- Grand challenge programs such as in Discovery Systems or Exploration will demand research in new areas.



Sincere Acknowledgements

- Dr. Barney Pell
- Dr. Robert Raskin, JPL
- Dr. Richard Keller
- Dr. David Maluf
- Dr. David Bell
- Dr. Daniel Berrios
- Ms. Jayne Dutra
- Mr. Everett Cary
- Dr. Gary Davis